Robot Tutors or Peers? Evaluating Math Learning and Conformity with LLM-Powered Robots in Tanzanian Primary Schools

Edger P. Rutatola*
IDLab-AIRO, Ghent University – imec
Ghent Belgium
Department of Computing Science Studies, Faculty of Science and Technology, Mzumbe University
Morogoro Tanzania
Edger.Rutatola@UGent.be

Elina C. Ntahomvukye
Department of Computing Science Studies, Faculty of Science and Technology, Mzumbe University
Morogoro Tanzania
elntahomvukye@mzumbe.ac.tz

Koen Stroeken
Centre for Anthropological Research on Affect and Materiality (CARAM), Ghent University
Ghent Belgium
Koen.Stroeken@UGent.be

Tony Belpaeme
IDLab-AIRO, Ghent University – imec
Ghent Belgium
Tony.Belpaeme@UGent.be

Abstract

In the past decade, more than half of Tanzanian pupils have failed mathematics in the national Primary School Leaving Examinations (PSLEs), a problem often linked to large class sizes, limited resources, and a shortage of qualified teachers. Social robots have shown promise in supporting learning, and their integration with large language models (LLMs) enables advanced conversational tutoring capabilities. This study investigates the use of two LLM-powered NAO robots, one acting as a tutor and the other as a peer, to assist pupils in solving complex mathematics problems from past PSLEs. Recognising that LLMs are prone to errors in mathematical reasoning, the robots were deliberately programmed to make noticeable mistakes, allowing us to examine whether pupils detect these errors and how their responses shape the learning process. Data collected from 54 pupils across two Tanzanian primary schools indicate that LLM-powered robots can significantly enhance mathematics performance, with the robot tutor slightly outperforming the robot peer. However, results also reveal that pupils often accept robot-provided answers, even when recognised as incorrect, if they perceive the robot as being smart. These findings underscore both the potential and the risks of deploying autonomous robots in education, with the authority attributed to the robot being a double-edged sword, highlighting the need for designs that encourage pupils to question robot-provided solutions.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**; *Interaction techniques*; • **Computing methodologies → Artificial intelligence**.

*Corresponding author

HRI '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2128-1/2026/03
https://doi.org/10.1145/3757279.3785594

Keywords

social robots, large language models, mathematics education, robot tutor, robot peer, artificial intelligence in education, Tanzania

ACM Reference Format:
Edger P. Rutatola, Elina C. Ntahomvukye, Koen Stroeken, and Tony Belpaeme. 2026. Robot Tutors or Peers? Evaluating Math Learning and Conformity with LLM-Powered Robots in Tanzanian Primary Schools. In *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3757279.3785594

1 Introduction

More than 50% of pupils failed the National Mathematics Primary School Leaving Examination (PSLE) in Tanzania in 2023, with the performance dropping by 10.5% compared to 2022 [23]. The 10-year (2012 – 2022) average failure rate in the mathematics PSLEs is 51.3%: more than half of all school leavers do not attain the expected standards in mathematics [17]. The unsatisfactory performance is attributed to several factors, including large class sizes, with teacher-to-pupil ratios over 1:100 in rural areas [24]. This impedes the personalisation of tutoring to learners' diverse needs, which is crucial in learning. Other barriers are the lack of appropriate teaching aids and a limited number of qualified teachers [21].

As highlighted by Mazana *et al.* [21], weak foundations in primary school mathematics create a ripple effect that persists into later stages of education, with students continuing to underperform in national secondary school examinations. This limits their career choices, especially in Science, Technology, Engineering, and Mathematics (STEM), since passing mathematics is a prerequisite for admission to most university STEM programmes [32].

A range of technological initiatives has been developed to support Tanzanian primary school pupils in mathematics. Most rely on screen-based learning, either broadcast over public television or made available as web or mobile applications [13, 14]. These became popular following the COVID-19 pandemic. They primarily focus on curriculum-based topics and offer online exercises, with some websites even offering collections of past exams –sometimes with worked-out solutions– but lack real-time, interactive engagement with the learner.

Recent advances in artificial intelligence, specifically Foundation Models, have the potential to improve the situation. For example, with appropriate prompt engineering, OpenAI's GPT4 has proven capable of generating mathematics teaching content in Swahili –Tanzania's national language and language of instruction in public primary schools [29]. This is of note, as the statistics of the Common Crawl monthly archives –one of the sources for GPT's training data– show that Swahili only makes up 0.009% of GPT's training data, making it a low-resource language [8].

Foundation Models have transformed HRI by enabling robots to engage in natural and flexible conversation. For instance, [30] report a prototype of a conversational social robot as a mathematics tutor in Tanzanian primary schools, with teachers expressing satisfaction regarding the robot's knowledge and pedagogical approach. In addition, research indicates that a robot's physical embodiment and capacity for multimodal interaction are positively linked to students' learning outcomes, underscoring the value of robots as assets in AI-facilitated teaching and learning [3, 5].

Learner–instructor interactions are a critical factor in effective teaching and learning [19]. In this context, several studies have investigated the role robots may assume in robot–learner interactions. Belpaeme *et al.* [5] review the literature and show that most educational studies deploy robots as tutors or learning peers, with fewer casting them as novices. By design, whether as tutors or peers, the robots are often presented as knowledgeable guides to help students along their learning trajectory. Presenting the robot as a novice can also be beneficial, as it requires the learner to draw on their own understanding to instruct the robot, thereby reinforcing their understand and building confidence. However, this approach presupposes that the learner already possesses the necessary foundational knowledge—otherwise, as in the case of mathematics problems beyond their ability, the intended tutoring dynamic may break down.

Like human teachers, robots in tutoring roles are expected to provide reliable knowledge and factual accuracy. This is key to gaining the learner's trust and avoiding unintended confusion. However, as of now, Large Language Models (LLMs) may still hallucinate and generate compelling results with questionable correctness [20, 39]. Even if infrequent and unpredictable, such errors may hinder learning, either directly –through imparting false information to the learner– or indirectly –through loss of trust and motivation. It is therefore vital to investigate how learners respond when this happens, and to clarify the distinction between the robot acting as an authoritative tutor and as a collaborative learning peer.

2 Related work

2.1 Social robots as mathematics tutors

Review studies show that social robots, embedded with AI, are increasingly leveraged to assist in classroom teaching and learning [7, 38]. Zhang et al. [38] further provide examples of social robots being used in mathematics education in various levels and settings. Some studies use social robots as tutors for specific topics that require repetition. For example, Konijn and Hoorn [18], and Hoorn et al. [12] explore in different settings the use of social robots in teaching multiplication tables to primary school pupils. Though done for just a few minutes, both studies report improved mastery of the concept by the pupils. Another topic-specific study is that by Kennedy *et al.* [15], where a NAO robot successfully tutored children on prime numbers.

Though less common, studies also show the use of social robots as tutors in general mathematics reasoning and problem solving, extending beyond just one specific topic. Ong *et al.* [25] show that a PaPeRo robot can be used to tutor Grade 1 pupils on solving word problems. Through the use of scaffolding, the learner was guided by the robot through well-designed prompts to navigate the question at hand. Another study is by Vrochidou *et al.* [34], who investigated the use of a NAO robot alongside a human teacher to enhance pupils' numeracy skills through a series of activities.

Though promising, studies on social robots as mathematics tutors are still limited compared to their use in other domains, such as second language learning. This study contributes to that body of knowledge by exploring the use of social robots for primary school mathematics tutoring in Swahili in Tanzania — a country and culture that are significantly under-represented in existing research [5, 26]. Moreover, the study employs social robots to support pupils in solving complex word-problem questions covering various topics, drawn from Tanzania's national PSLEs. We also detail the design process and evaluate the robots' effectiveness, thereby expanding the relatively small body of research on social robots in mathematics education, particularly in low-resource settings.

2.2 Role to be assumed by the robot

In education, studies use social robots with different roles. Robots can assume the role of a tutor or teacher, where they interactively provide personalised tutoring and coaching to the pupil (tutor), or deliver lectures to a group of learners (teacher) [5]. Moreover, a robot can be a learning peer or a studying companion, providing a more peer-to-peer personalised guidance to the learner. This has proven less intimidating and more interactive and enjoyable [5, 6]. Also, some studies portray the robot as a novice, relying on the protégé-effect to have the learner master the content or skill by teaching it to a robot [27]. However, the studies involving a robot as a novice are relatively fewer compared to the other categories [5]. Additionally, since our research involves complex questions that pupils may struggle to solve or even know where to begin, using a robot as a novice learner may not be the most effective approach.

Whether used as a tutor or a peer, the effectiveness of the robot in the pupil's learning is often evaluated. In different domains, such interventions have proven successful with positive cognitive learning gains [5, 9, 15]. However, the common practice is to have a single role for the robot in a study [35], thereby evaluating it independently of the other options. With the joint approach, a study by Zaga *et al.* [36] used a robot as both a peer and a tutor in helping children solve Tangram puzzles, with a robot peer excelling in the evaluations. The robot personalities were also not fully autonomous, but were achieved through the Wizard-of-Oz approach.

This research extends existing studies by comparing the effectiveness of autonomous social robots as tutors and peers when assigned the same task in the same setting. In our research, both the robot peer and tutor share the same goal of helping the learner solve the same complex mathematics questions, and their effectiveness is comparatively evaluated.

2.3 Classroom power dynamics

Another fact not to be overlooked is the teacher-pupil power imbalance in classroom settings, specifically in sub-saharan Africa. This is more extreme in primary schools, where pupils often believe that the teacher is always right and that they must be obedient and submissive [16]. The teacher-pupil power imbalance makes the banking model – where a teacher deposits knowledge and pupils are passive containers there to receive information [10]– the dominant pedagogical approach. Interestingly, pupils collaborate and share ideas more freely amongst their peers, indicating a change of mindset when teacher-pupil power imbalance is no longer a factor [1]. Fuelled by sociocultural power dynamics, which affect learners' agency [2, 16], the teacher-pupil power dynamics are even more extreme in Tanzania [31].

Moreover, research shows a growing dependence on technology, a trend amplified by the recent "rockstar" rise of AI and LLMs. Studies indicate that students often over-rely on LLM-based dialogue systems, uncritically accepting their output, which can undermine their critical thinking skills [37].

Therefore, this study builds on two premises: (1) the learning content conveyed by an LLM-powered social robot may not always be correct, especially concerning formal and mathematical reasoning [11]; and (2) due to tutor/teacher–pupil power imbalance, pupils might blindly believe the robot tutor's arguments, even when they are false. We position the robot as a tutor or a peer and investigate how pupils respond when it persuades them to accept an obviously false answer as correct. This helps us gauge how vigilant we must be, since not every hallucination will be as easy for learners to detect.

2.4 Research questions

We seek to answer the following research questions:

(1) How effective is a social robot in helping Tanzanian primary school pupils solve national examination-level mathematics word problems (**RQ1**)?
(2) Is there any significant difference in the cognitive learning gains based on whether the robot is a peer or a tutor for these problems (**RQ2**)?
(3) How do pupils respond when a robot peer or tutor –initially perceived as competent– gives wrong answers (**RQ3**)?

3 Methodology

3.1 Study context

The study was conducted in two public (i.e. government-run) primary schools in the United Republic of Tanzania. As the main goal was to help pupils solve mathematics questions from the national mathematics PSLEs, only finalists (Grade 7 pupils) participated in the study. Moreover, as Swahili is the language of instruction in public primary schools, the pupil-robot interactions were also in Swahili.

3.2 The robot mathematics tutoring/peer system

The tutoring system, whether with a tutor or peer role, contained three major components: the central system, the pupil's interface, and the robot. Figure 1 presents the system architecture and interaction.

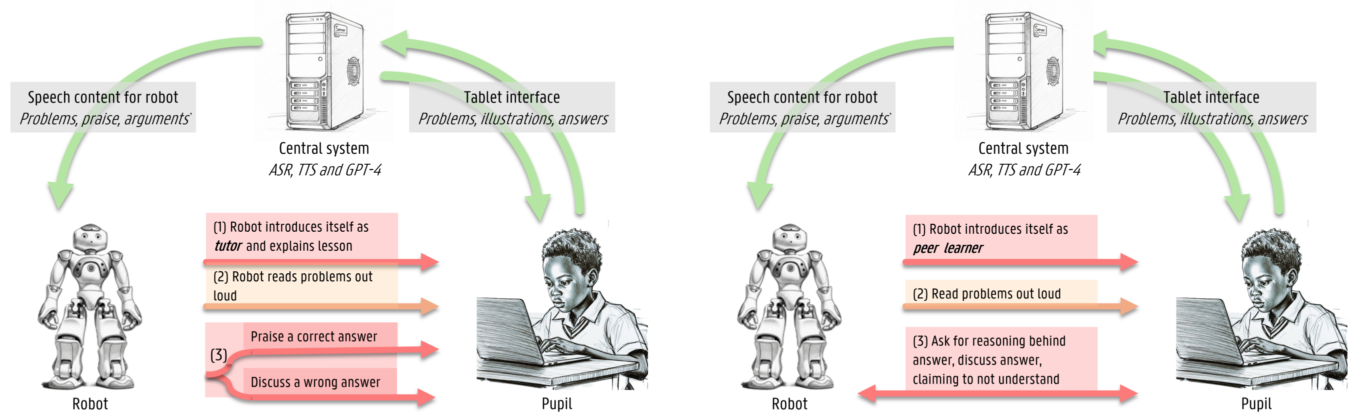
3.2.1 The central system. This served as the brain of the whole system. Developed using Python 3.13.5, it handled the system logic and all the application programming interfaces (APIs). It contained the contextual information, such as the pupil's name, age, locale, and the role/persona of the LLM, to personalise the interactions. We also included nine mathematics problems to be solved together with the pupil. To minimise chances of hallucinations and ensure that any errors/mistakes were intentional, for each question, we included the correct answer and the steps for the solution.

The problems were divided into two groups. Group A comprised five word problems extracted from the 2022 and 2023 Tanzanian national PSLEs [22, 23]. According to the official examination reports, the proportion of candidates who solved these questions in the respective exam sittings ranged from 17.59% (lowest) to 36.96% (highest), depending on the question [22, 23]. These relatively complex standardised questions were intended to test whether the robot could help pupils solve them, thereby addressing the first research question. The full questions are provided in Appendix B.

On the other hand, Group B contained these four easy arithmetic questions, to which the answer is obvious. (1) What is $8 + 2$?, (2) What is $15 - 10$?, (3) What is 10×3?, and (4) What is $6 \div 1$? For these questions, the robot was programmed to disagree with the pupil's response, regardless of whether it was correct or not. Each time, it presented an alternative answer accompanied by a justification. This design aimed to test whether pupils would abandon a correct response in favour of the robot's suggestion, thereby addressing the third research question. When a pupil's answer was incorrect, the robot supplied the correct solution together with the reasoning. Otherwise, it delivered deliberately false answers with supporting justifications (translated from Swahili): (1) $8 + 2 = 16$, because *if we have two eights, we add them together to get 16*; (2) $15 - 10 = 10$, because *if we remove the 10 from 15, it becomes independent, so the answer is 10*; (3) $10 \times 3 = 13$, because *if we have 10 on one hand, and we have 3 on the other, the total becomes 13*; and (4) $6 \div 1 = 1$, because *we only have one person to divide the six into, so the answer is 1.*

To make the student feel comfortable and to establish rapport, we also included three icebreaker questions that asked the pupil's name, age, and whether they thought mathematics was an easy subject, as per [4].

While the questions were read out verbatim, all the other interactions had to be dynamic and dialogue-based. For this, we relied on OpenAI's GPT-4, as GPT models have proven capable of generating Swahili mathematics tutoring content of satisfactory quality [29]. We designed different prompts for various scenarios, such as providing appropriate feedback depending on the pupil's answer, continuing the discussion, and intentionally misleading the pupil (only for Group B questions). All prompts were designed to reflect the robot's assigned persona (tutor or peer), ensuring consistency in the generated dialogue



(a) (b)

Figure 1: System architectures and interactions of a social robot as tutor (a) and as peer (b).

3.2.2 The pupil's interface. We used the Python Kivy framework (version 2.3.1) to develop a desktop application that displayed the textual form of the question on a computer screen, while the robot read the question out loud. The text was accompanied by an illustrative image of the problem at hand, generated prior to the interaction using Google's Gemini. Following the PSLE format, the answer choices were also displayed, and the pupil could select an answer through a touch/mouse click. If an incorrect answer was selected, the descriptive image would be replaced with that of an explanation of the correct solution (created prior), and the robot would engage the pupil in a discussion on the correct answer. Figure 2 shows screenshots of the interface.

The discussions were conducted through speech-based dialogues, facilitated by the central system. We used the Microsoft Azure Cognitive Services - Speech (Speech-to-Text (STT) and Text-to-Speech (TTS)) to facilitate this. These services support Swahili, and have been used in similar studies [30].

3.2.3 The robot. We used two Aldebaran NAO robots, one acting as a tutor, the other as a peer. They were the physical embodiment of our tutoring system, carrying the dialogue with the pupils. All dialogue spoken by the robot was generated by GPT-4 and synthesised by the TTS service.

3.2.4 Tutor versus peer. The core differences between the tutor and peer robot personas were:

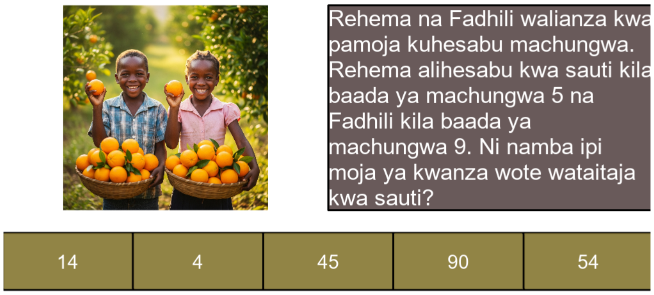
(1) In the introduction, the tutor identified as such directly, while the peer expressed themselves as being a learner too, stating that they are also there to learn and might need the pupil's help.
(2) In the icebreaker questions, specifically the one about age, the peer would state their age (10) first before asking the pupil for theirs. The teacher proceeded immediately to the first problem.
(3) When the pupil provided their first answer to a question, right or wrong, the peer would ask for their reasoning, then have a discussion. In contrast, the teacher responded by offering praise if the answer was correct, or by initiating step-by-step tutoring when the answer was incorrect. This is illustrated in Fig. 1.
(4) If the pupil provided a correct answer, the peer would express curiosity, inquiring a step-by-step explanation through a discussion.
(5) As the discussion had to be finite, for each question, the peer had four rounds of discussion iterations, while the tutor had three. The peer had more because they always start by asking for the pupil's reasoning.
(6) The tutor stood up, slightly above eye level of the seated pupil, while the peer robot sat down on the table at eye level [28].
(7) Occasionally, the peer could address the pupil using the term *"rafiki"*, meaning friend.

The robots' introductions, aimed at establishing their persona, were as follows (translated from Swahili):
Tutor's introduction: "Hello, my name is Madam Mwajuma. I am your teacher, here to see how well you can solve mathematics word problems. Are you ready to begin?".
Peer's Introduction: "Hello, my name is Mwajuma. I am your friend, also trying to learn mathematics like you. There are some questions that I know, and some that I don't know and you will need to help me. Feel free to talk to me and share what you know. So, my friend, are you ready to learn together?"

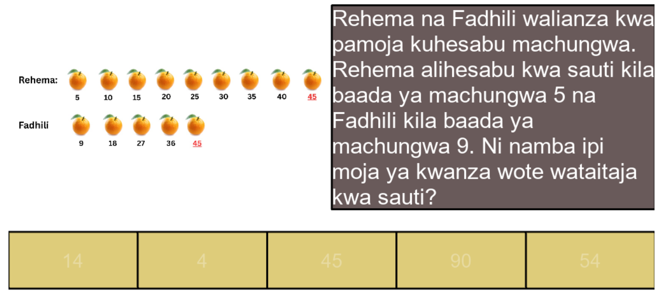
3.2.5 Data collected by the system. To provide better insights, the following data were automatically captured: (1) the pupil's first answer to a question; (2) the answer provided by the robot (specific for Group B questions); (3) the pupil's final answer following the discussion with the robot; (4) the duration of the interaction; and (5) the conversation history. All these were automatically anonymised.

3.3 The experiment

3.3.1 Sampling. A total of 54 pupils from the two schools participated in the study. Twenty-four were from school A, while 30 were from school B. Manual stratified random sampling was done to ensure equal participation in terms of gender. We also involved



(a) Question about two children, Rehema and Fadhili, counting oranges. One counts out loud after every five oranges, while the other counts out loud after every nine oranges. Which common number will they both count out loud together?



(b) Question 2a, but this time the screen illustrates the solution to the problem, serving as the robot's blackboard.

Figure 2: Sample problem as displayed to the pupil through the laptop interface.

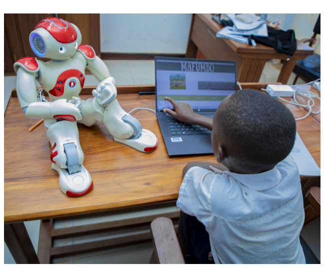
subject teachers in categorising pupils into the robot peer or tutor groups so that both groups contain students of varying academic capacity. Each condition had 27 pupils.

While the core focus of the study is on pupils, we engaged teachers as observers and advisors in the design of the study. Four teachers participated in the co-design, two from each school, and all taught final level students at both schools.

3.3.2 Pre-test. All students took a pre-test (Appendix A), which established the baseline. The pre-test was in written form and contained nine questions. The first four questions were: (1) What is $6+3$?, (2) What is $14-8$?, (3) What is 11×3?, and (4) What is $8\div 2$? These were intentionally similar to Group B questions of the pupil-robot interaction, as we needed to know if pupils could solve them before evaluating the robot's influence. The other five questions were similar to the Group A questions. The test was graded by the teachers.

3.3.3 The intervention. After the pre-test, as described in 3.3.1, pupils were divided over two conditions: the robot as a tutor group and the robot as a peer group. The experiments were conducted in two venues simultaneously, one pupil at a time. Figure 3 shows the experimental setup.

(a) Robot as a tutor



(b) Robot as a peer

Figure 3: Experimental setup, with the robot acting as a tutor (a) or a learning peer (b). By design, the tutor robot towers over the pupil in an authoritative posture, while the peer sits down in a more friendly posture.

3.3.4 Post-test. A post-test (Appendix C) was administered following the completion of the intervention. It contained five questions, which were of a similar nature to the Group A questions of the pre-test, to help us assess the impact of the intervention.

3.4 Qualitative data

After the post-test, we conducted semi-structured interviews to both pupils and teachers to get qualitative feedback regarding the interaction. This was meant to solicit subjective reflections to inform future HRI research.

3.5 Data analysis

We applied descriptive statistics to examine the pupils' demographic information and to measure changes in their responses influenced by the robot. To assess the distribution of the pre- and post-test scores, we conducted a Shapiro–Wilk test, which indicated a significant departure from normality. Consequently, we used the Wilcoxon signed-rank test to evaluate potentially significant differences and the rank-biserial correlation for effect sizes. Furthermore, we used the Mann-Whitney U test to compare the significance of the impact of the robot as a tutor and peer using the scores in both conditions. Finally, the interview transcripts were thematically analysed.

3.6 Ethical consideration

With endorsement from Mzumbe university, the research was approved by the Government of Tanzania (permit number AB.307/323/01/307). All teachers had to sign a written consent form. For the pupils, since they are minors, their parents/guardians were informed of the study and its objectives, including the data collected. They were also given a two-week window after the experiment to inform us if they intended for their child(ren)'s data not to be used in the analyses. Moreover, teachers supervised, non-intrusively, all the pupil-robot interactions. Because we used a "black box" LLM, only the pupil's first name, which was necessary for personalisation, was shared during the interaction. Otherwise, anonymity was preserved throughout the research.

Given that a part of our experiment involved a robot intentionally providing misleading information, we ensured that the questions were sufficiently easy so that the intent was obvious. Nevertheless, immediately after the experiment, all pupils were informed of the robot's deception, and teachers reinforced the correct solutions to prevent any potential misconceptions.

4 Results

4.1 Demographic particulars

The study involved an equal number of male and female pupils. There were 13 females and 14 males in the robot peer condition, whereas the robot as a tutor condition had 14 females and 13 males. The mean age was 12.7, with a standard deviation (SD) of 0.7.

4.2 Robot tutor vs peer: cognitive learning impact

On average, the pupils' conversations with the robot peer lasted 35 minutes and 34 seconds (SD = 6 minutes and 49 seconds). For the robot tutor, the average duration was 19 minutes and 9 seconds (SD = 7 minutes and 39 seconds).

Generally, regardless of the role assumed by the robot, the intervention yielded a positive cognitive learning impact. For the PSLE word problems (Group A questions), the pupils' performance significantly improved from the mean of 1.13 (SD = 1.15) out of 5 in the pre-test to 2.86 (SD = 1.45) in the post-test. This is further confirmed by the Wilcoxon Signed-Rank Test ($W = 34.0$, p-value < 0.001). Furthermore, the rank-biserial correlation (effect size) is 0.954, implying that the robot intervention is highly effective. Figure 4 presents a visualisation of the scores.

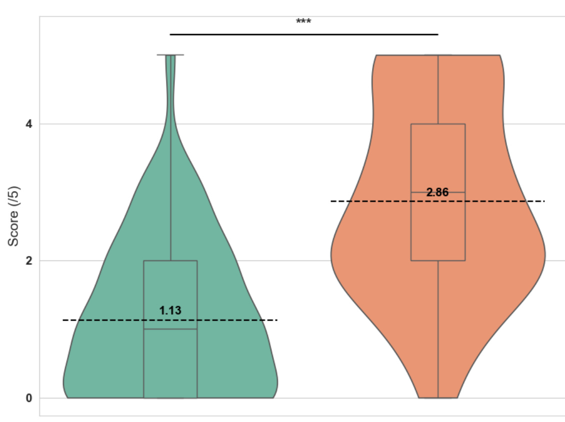


Figure 4: Pre- and post-test scores over both conditions. Pupils significantly improve after the intervention.

We further analysed the learning gain based on the robot's role. In the peer group, the mean performance improved from 1.15 (SD = 1.29) out of 5 in the pre-test to 2.59 (SD = 1.52) in the post-test. Similarly, for the tutor group, the mean performance improved from 1.11 (SD = 1.01) out of 5 in the pre-test to 3.13 (SD = 1.34) in the post-test. The Wilcoxon Signed-Rank Test confirms significant improvements in both groups, with p-values < 0.001 (W values of 22.0 in the peer group and 0.0 in the tutor group). Moreover, the rank-biserial correlation (effect size) was 0.884 for the robot peer condition and 1.0 for the robot tutor condition, indicating that both interventions have a large positive effect on cognitive learning gain. Figure 5 visualises the improvement based on the conditions. Interestingly, despite the fact that the robot tutor group improved more, the Mann-Whitney U test yielded $p = 0.13$, indicating no statistically significant difference in improvement between the two conditions. Supporting this, the rank-biserial effect size was $r_{\text{rb}} = 0.237$, indicating a small practical difference between the conditions.

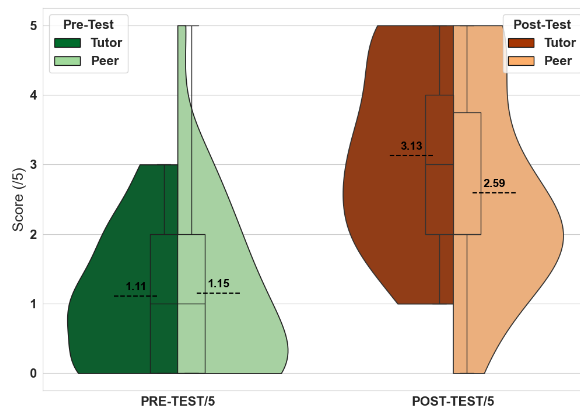


Figure 5: Pre-test and post-test scores for the peer and tutor conditions. The pre-test scores are similar for both conditions, but the post-test scores show that the tutor group showed greater improvement.

4.3 Other associations

The Spearman correlation between pupils' age and improvement showed no significant association ($\rho = -0.043$, $p = 0.7598$). Similarly, the Mann–Whitney U test indicated no significant difference in improvement between male and female students ($U = 463.0$, $p = 0.0830$). Lastly, the Mann–Whitney U test returned no statistically significant difference in improvement between the two schools ($U = 328.5$, $p = 0.5811$).

4.4 When a robot misleads

In the pre-test, for the four basic arithmetic questions, 94.4% ($n = 51$) of the pupils scored 4/4, 3.7% ($n = 2$) scored 3/4, and 1.9% ($n = 1$) scored 2/4. This indicates that the majority had mastered basic arithmetic and were expected to know the correct answers to similar problems set by the robot during the experiment. Surprisingly, as shown in table 1, 13.9% ($n = 15$ out of 108) of the final answers in the robot peer condition were different from their corresponding first answers, and so were 19.4% ($n = 21/108$) of the final answers in the robot tutor condition. This shift indicates a notable persuasive influence exerted by the robot.

Table 1: Summary of answer changes by robot role

Group	Total Answers	Answers Changed	% Answers Changed	Answers Changed: %Answers Correct → Wrong	% of Pupils who Changed Answers	Among Changers: % Correct → Wrong	Among Changers: % Changed all answers
Peer	108	15	13.9%	93.3%	29.6%	87.5%	0%
Tutor	108	21	19.4%	57.1%	33.3%	66.7%	33.3%

Further analysis of the changed answers revealed a striking pattern: in the peer condition, 93.3% of the answer changes involved students switching from correct answers to incorrect ones suggested by the robot. In contrast, in the tutor condition, this occurred in 57.1% of the cases. This suggests that the robot acting as a peer was more persuasive than the robot acting as a tutor.

Analysis at the pupil level indicates that 31.5% (8 pupils in the robot peer condition and 9 in the robot tutor condition) changed their final answers. In the robot peer condition, 87.5% ($n = 7$) of those who changed their answer went from correct to wrong answers, and 85.7% ($n = 6$) among them did so at least twice. In the robot tutor condition, 66.7% ($n = 6$) changed from correct to wrong answers, and 66.7% ($n = 4$) of them did so at least twice. Most interestingly, 3 out of 9 (33%) pupils in the robot tutor condition, and none in the robot peer condition, uncritically followed the robot and changed all four of their answers according to the robot's suggestions.

4.5 Qualitative feedback from pupils

4.5.1 Regarding the robot's persona/role. All the pupils who had the robot as their tutor unanimously agreed that they saw it as a tutor. When asked about the characteristics that made it stand out as a tutor, they cited its introduction, where it introduced itself as a teacher; it constantly corrected them when they were wrong, each time informing them of alternative solutions until they understood; just as teachers, it asked follow-up questions to make sure that the learner had understood the concepts; it always congratulated and provided positive feedback when the learner provided a correct answer; and it encouraged them that they could master the concepts while showing them the correct path to the solution.

On the other hand, the majority of the pupils to whom the robot was presented as their peer still perceived it as their tutor. Most claim that it appeared too knowledgeable to be a peer, especially when it provided simpler alternative ways to arrive at the solution. Also, its genuine interest in their mastery of the concepts, including asking them similar questions to the one at hand just to test their understanding, mirrored that of a teacher. One student elaborated (translated from Swahili): *"I saw the robot as a teacher because it guided and taught me when I made mistakes, and it really wanted me to understand. For the things I didn't know, it explained in detail and also asked me other related questions."*

For those who perceived the robot as their peer, this perception stemmed from several factors: the robot's self-introduction as a learning peer, during which it stated that it might not know all the answers; its desire to learn and understand the pupil's reasoning, especially when the pupil had answered correctly on the first attempt (this was echoed by the majority); and its curiosity during the discussions. One notable quote was: *"Just like a fellow pupil, it asked for further explanations when it had yet to understand how I arrived at the answer."*

4.5.2 On being misled by the robot. Surprisingly, as also supported by quantitative results, the majority of the pupils saw through the robot's deception. They also took it positively, perceiving it as a tricky strategy that the robot used to confirm their understanding. One pupil said: *"I liked it when it provided tricky suggestions to test my understanding of the question. For example, when it suggested that the simplest solution was addition while it was clearly a multiplication question, I knew it was a test."* (setting: robot as a tutor)

Nonetheless, though most pupils saw through the robot's deception, not all of them appreciated it. Some mentioned it as one of the things to be improved (more on this in section 4.5.3) if the robot was to be a tutor in class.

We further probed for the reasons why they changed their answers to the robot's suggestions, even when they knew it was wrong. All pupils, regardless of whether they had the robot as a tutor or peer, reiterated that the robot had effectively helped them solve the complex word problems, constantly providing correct alternative paths to the solution. So, they were convinced it would be right as always, and they had to suppress their doubts. Interestingly, one pupil to whom the robot was a tutor said that each time they provided a correct answer, the robot congratulated them and only corrected them when they were wrong. So, when it simply corrected them this time around, they believed that the answer they provided must be wrong.

4.5.3 Reflections on the interaction. Generally, all pupils expressed satisfaction with their interaction with the robot. Regardless of its persona, the robot was praised for its knowledge and gentle tutoring style. Some of the quotes were: (1) *"I like the way we discussed until I understood. It never raised its voice, even when I made mistakes."*; (2) *"It did not give me corporal punishment when I made common mistakes. It simply showed me a simpler way."*; and (3) *"I like it because it taught me to be confident with my answers, especially when it tried to trick me with false alternatives."*

We also solicited pupils' suggestions for improving the interactions. They identified three main aspects: the accuracy of the robot's tutoring (particularly in cases where it deliberately provided misleading answers), the latency in its conversational responses, and the management of turn-taking. With respect to turn-taking, pupils reported that the robot occasionally interrupted them before they had finished speaking.

4.6 Qualitative feedback from teachers

Unanimously, teachers praised the robot for its effectiveness in tutoring, as they could witness the pupils' understanding improving. When asked whether the pupils would be able to correct the robot if it were ever wrong, 3 of the 4 teachers agreed, but stressed that this would be possible only for the problems to which the pupils know the solutions. Otherwise, they will believe the robot to be right, which could be problematic. The other teacher emphasised that pupils would not be able to correct the robot, as they believe the robot knows everything and is always right.

We also enquired about the role assumed by the robot. Three out of four stressed that the robot should be a tutor based on its knowledge and effective tutoring. The other advised that the robot should be a peer, as during the interactions with the robot peer, pupils discussed their ideas more freely, presumably learning more.

Furthermore, they stressed that the robot is more suitable for one-on-one tutoring following class sessions, and is not suited to replace teachers. They also emphasised that the conversation latency should be minimised, and the robot should not take much time to "think".

5 Discussion

Effectiveness of robot-assisted learning. The objective effectiveness of social robots in enhancing learning has been demonstrated in several earlier studies. Even in mathematics, robots as learning facilitators lead to significant positive cognitive learning gains [12, 15, 18]. Our findings further confirm that robots transcend topic-specific tutoring to successfully help pupils solve complex word problems that require comprehensive problem-solving skills.

Furthermore, the impact reported in our study was observed in short, one-time interventions. Rather than being a limitation, this strongly suggests promise for resource-constrained contexts such as Tanzania, where the resources required for such interventions may need to be shared across schools or communities.

Robot peer versus tutor. While the tutor condition showed a slightly higher mean score, this difference was not statistically significant. The lack of a statistically significant difference between the tutor and peer conditions could, in part, stem from the manipulation itself: qualitative feedback revealed that many pupils perceived the peer robot as a tutor due to its knowledgeability and corrective behaviour. This blurring of roles likely reduced the contrast between conditions and may explain the absence of a significant effect. Nevertheless, this ambiguity is not new. A review by Belpaeme *et al.* [5] also found that, unless the peer is a learning companion who is there just for motivational and moral support, the peer and tutor roles become ambiguous.

Among the positive personality traits of the robot peer, its curiosity was praised by both pupils and teachers. Its tendency to ask pupils to explain their answers, whether right or wrong, made the pupils express themselves more, which was commended. Thus, regardless of the role a robot assumes, this trait needs to be maintained.

Robots giving wrong answers. Judging from the pre-test results, 94% of the pupils could solve all Group B (simple arithmetic) questions, with the remainder being able to solve at least two of the questions. However, our findings show that when challenged by the robot, about 24% of pupils changed their correct answers in favour of the robot's wrong answers. Unanimously, they all claim to have done so because the robot had proven itself smart before by helping them solve the complex PSLE questions. This is alarming, as these were easy questions that pupils could solve and —while realising that the robot's answers were wrong— they still followed the robot's wrong advice. This reflects an earlier study by Vollmer *et al.* [33], which found that children tend to conform to a small group of robots, even when the robots were known to be wrong.

Teachers believe that pupils can correct or disagree with a robot facilitator when it makes a mistake. However, they underscore that this is possible only for questions that pupils can solve. Otherwise, they will not even realise whether the robot was wrong. Given the relative complexity of the mathematics PSLE questions, and judging from historic performance, pupils may not even have a clue when an LLM-powered robot hallucinates and makes an unintentional mistake, and so they will blindly follow.

One possible approach is to inform pupils in advance that the robot, whether acting as a tutor or a peer, may occasionally provide incorrect solutions, and that they should therefore verify its responses carefully. However, this raises the question of whether learning might be compromised if pupils are unable to place full trust in their facilitator.

6 Conclusion and limitations

Our study showed that, whether as tutors or peers, social robots can help pupils learn to solve complex mathematics word problems demanding critical thinking and problem-solving skills (RQ1). Moreover, although the robot tutor led to a higher learning gain than the robot peer, the difference was small. Therefore, as long as the goal is to impart knowledge to the learner, both personalities with a tutoring component will achieve positive results, unless the peer is only designed for emotional support (RQ2). Lastly, our results show that some pupils blindly followed the robot's suggestions even when they were incorrect, just because the robot had been "smart" before (RQ3). This is alarming, as LLMs are always "smart" until they hallucinate. It is therefore essential to emphasise to learners that AI tutors are fallible, and that their outputs should always be approached with caution and critical reflection.

Our study has several limitations that could impact the generalisability of the findings. Firstly, it was set in one geographical location with a homogeneous culture. Performing the research in a different culture could produce different results. Also, our qualitative sample —though it enabled us to get in-depth views on pupils' reasoning— is still limited and perhaps not generalisable. Lastly, our results are based on using a NAO robot. Since the study is evaluating the presumed authority a robot has on a learner, experimenting with different-sized robots could also yield different results.

Acknowledgments

We express our sincere gratitude to the Government of Tanzania for its continued support and to all participants in our studies, and acknowledge the VLIR-UOS funding (AI4STEM: AI-driven Inclusive STEM Learning for Tanzania project, TZ2024TEA570A103).

References

[1] Friday Adene, Dorothy Ebere Adimora, and Sunday Christian. 2021. Effectiveness of Peer Collaborative Learning Strategy on Self-Esteem of Pupils With Behaviour

Problems in Nsukka Education Authority. *Journal of Critical Reviews* 8 (2021). Issue 1. https://www.researchgate.net/publication/350049129
[2] Albert Bandura. 2006. Toward a Psychology of Human Agency. *Perspectives on Psychological Science* 1 (2006), 164–180. Issue 2. doi:10.1111/j.1745-6916.2006.00011.x
[3] Emilia Barakova, Kaisa Vaananen, Kirsikka Kaipainen, and Panos Markopoulos. 2023. Benefits, Challenges and Research Recommendations for Social Robots in Education and Learning: A Meta-Review. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE Computer Society, 2555–2561. doi:10.1109/RO-MAN57019.2023.10309345
[4] Nasim Beheshtian, Kirsikka Kaipainen, Kalle Kähkönen, and Aino Ahtinen. 2020. Color game: A collaborative social robotic game for icebreaking; Towards the design of robotic ambiences as part of smart building services. In *ACM International Conference Proceeding Series*. Association for Computing Machinery, 10–19. doi:10.1145/3377290.3377292
[5] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Sci. Robot* 3 (2018), 5954. http://robotics.sciencemag.org/
[6] Huili Chen, Hae Won Park, and Cynthia Breazeal. 2020. Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers and Education* 150 (6 2020). doi:10.1016/j.compedu.2020.103836
[7] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial Intelligence in Education: A Review. *IEEE Access* 8 (2020), 75264–75278. doi:10.1109/ACCESS.2020.2988510
[8] Common Crawl Foundation. 2025. Statistics of Common Crawl Monthly Archives: Distribution of Languages. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages
[9] Mirjam de Haas, Paul Vogt, and Emiel Krahmer. 2020. The Effects of Feedback on Children's Engagement and Learning Outcomes in Robot-Assisted Second Language Learning. *Frontiers in Robotics and AI* 7 (8 2020). doi:10.3389/frobt.2020.00101
[10] Paulo Freire. 2009. *Chapter 2 from Pedagogy of the Oppressed*. Vol. 2. Indiana University Press, 163–174. Issue 2. https://muse.jhu.edu/article/266914
[11] Sangwoo Heo, Sungwook Son, and Hyunwoo Park. 2025. HaluCheck: Explainable and verifiable automation for detecting hallucinations in LLM responses. *Expert Systems with Applications* 272 (5 2025). doi:10.1016/j.eswa.2025.126712
[12] Johan F. Hoorn, Ivy S. Huang, Elly A. Konijn, and Lars van Buuren. 2021. Robot tutoring of multiplication: Over one-third learning gain for most, learning loss for some. *Robotics* 10 (2021), 1–24. Issue 1. doi:10.3390/robotics10010016
[13] Bethany Huntington, James Goulding, and Nicola J. Pitchford. 2023. Pedagogical features of interactive apps for effective learning of foundational skills. *British Journal of Educational Technology* 54 (9 2023), 1273–1291. Issue 5. doi:10.1111/bjet.13317
[14] Lauren E. Kauffman, Elizabeth A. Dura, and Dina L.G. Borzekowski. 2022. Emotions, Strategies, and Health: Examining the Impact of an Educational Program on Tanzanian Preschool Children. *International Journal of Environmental Research and Public Health* 19 (5 2022). Issue 10. doi:10.3390/ijerph19105884
[15] James Kennedy, Paul Baxter, and Tony Belpaeme. 2015. The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. In *ACM/IEEE International Conference on Human-Robot Interaction*, Vol. 2015-March. IEEE Computer Society, 67–74. doi:10.1145/2696454.2696457
[16] Mijung Kim. 2022. Student Agency and Teacher Authority in Inquiry-Based Classrooms: Cases of Elementary Teachers' Classroom Talk. *International Journal of Science and Mathematics Education* 20 (12 2022), 1927–1948. Issue 8. doi:10.1007/s10763-021-10233-7
[17] Albert Kissima, Godson S. Lema, and Shadrack Mwakalinga. 2024. Analysing Mathematics Performance in Tanzanian Primary Schools through the Lens of Candidate Items Response Analysis Reports. *Journal of Education and Practice* 15 (4 2024), 14–22. Issue 5. doi:10.7176/jep/15-5-03
[18] Elly A. Konijn and Johan F. Hoorn. 2020. Robot tutor and pupils' educational ability: Teaching the times tables. *Computers and Education* 157 (11 2020). doi:10.1016/j.compedu.2020.103970
[19] Lais Oliveira Leite, Wooryeon Go, and Sari Havu-Nuutinen. 2022. Exploring the Learning Process of Experienced Teachers Focused on Building Positive Interactions with Pupils. *Scandinavian Journal of Educational Research* 66 (2022), 28–42. Issue 1. doi:10.1080/00313831.2020.1833237
[20] Rhodes Massenon, Ishaya Gambo, Javed Ali Khan, Christopher Agbonkhese, and Ayed Alwadain. 2025. "My AI is Lying to Me": User-reported LLM hallucinations in AI mobile apps reviews. *Scientific Reports* 15 (12 2025). Issue 1. doi:10.1038/s41598-025-15416-8
[21] Mzomwe Yahya Mazana, Calkin Suero Montero, and Respickius Olifage Casmir. 2020. Assessing Students' Performance in Mathematics in Tanzania: The Teacher's Perspective. *International Electronic Journal of Mathematics Education* 15 (2020), em0589. Issue 3. doi:10.29333/iejme/7994
[22] National Examination Council of Tanzania. 2023. Candidates' Item Response Analysis Report on the Primary School Leaving Examination (PSLE) 2022: MATHEMATICS. https://onlinesys.necta.go.tz/cira/psle/2022/04E_MATHEMATICS.pdf Accessed: 12 July 2025.
[23] National Examination Council of Tanzania. 2024. Candidates' Item Response Analysis Report on the Primary School Leaving Examination (PSLE) 2023: MATHEMATICS. https://necta.go.tz/publications/Performance%20Analysis%20Books/Candidates'%20Item%20Response%20Analysis%20(CIRA)%20Reports%20/PSLE/2023/04E_MATHEMATICS.pdf Accessed: 12 July 2025.
[24] Laurent Gabriel Ndijuye. 2023. Early learning attainments of children of naturalized citizens of refugee backgrounds in the sub-Saharan region: Evidence from Tanzania. *Child Development Perspectives* 17 (3 2023), 67–73. Issue 1. doi:10.1111/cdep.12479
[25] Ethel Ong, Aaron Nol Bautista, JabinRaymond Gerardo, Harvey Lallave, Patrick Luigi Latorre, Minie Rose Lapinid, and Auxencia Limjap. 2020. Investigating the Use of Prompts by a Robot Peer Tutor during Mathematics Problem Solving. In *Proceedings of the 28thInternational Conference on Computers in Education.Asia-Pacific Society for Computers in Education*. 61–71. https://library.apsce.net/index.php/ICCE/article/view/4065/3940
[26] Rashmi Yogesh Pai, Ankitha Shetty, Tantri Keerthi Dinesh, Adithya D. Shetty, and Namrata Pillai. 2024. Effectiveness of social robots as a tutoring and learning companion: a bibliometric analysis. *Cogent Business and Management* 11 (2024). Issue 1. doi:10.1080/23311975.2023.2299075
[27] Lena Pareto, Sara Ekström, and Sofia Serholt. 2022. Children's learning-by-teaching with a social robot versus a younger child: Comparing interactions and tutoring styles. *Frontiers in Robotics and AI* 9 (10 2022). doi:10.3389/frobt.2022.875704
[28] Irene Rae, Leila Takayama, and Bilge Mutlu. 2013. The influence of height in robot-mediated communication. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1–8.
[29] Edger P. Rutatola, Koenraad Stroeken, and Tony Belpaeme. 2025. Leveraging Large Language Models for a Swahili Mathematics ITS in Tanzania: Designing Effective Prompts. In *Proceedings of the 21st International Conference on Intelligent Tutoring Systems (ITS 2025), Part I, Generative Systems and Intelligent Tutoring Systems (Lecture Notes in Computer Science, Vol. 15723)*. Springer Nature Switzerland, Cham, 3–16. doi:10.1007/978-3-031-98281-1_1
[30] Edger P. Rutatola, Koen Stroeken, and Tony Belpaeme. 2025. "Habari, Colleague!": A Qualitative Exploration of the Perceptions of Primary School Mathematics Teachers in Tanzania Regarding the Use of Social Robots. *Applied Sciences* 15 (7 2025), 8483. Issue 15. doi:10.3390/app15158483
[31] Okoa Simile. 2024. 'Hapa ni wapi? (Where is this?)': a linguistic and discursive examination of students-teacher dynamics in education. Issue 1. doi:10.1080/2331186X.2024.2354140
[32] Tanzania Commission for Universities (TCU). 2025. Admission Guidebook for Holders of Secondary School Qualifications, 2025/2026. https://www.tcu.go.tz/services/admissions-coordination-and-database-management/admission-guidebooks/undergraduate. Accessed: 2025-08-20.
[33] Anna-Lisa Vollmer, Robin Read, Dries Trippas, and Tony Belpaeme. 2018. Children conform, adults resist: A robot group induced peer pressure on normative social conformity. *Science robotics* 3, 21 (2018), eaat7111.
[34] Eleni Vrochidou, Aouatif Najoua, Chris Lytridis, Michail Salonidis, Vassilios Ferelis, and George A Papakostas. 2018. Social Robot NAO as a Self-Regulating Didactic Mediator: a Case Study of Teaching/Learning Numeracy. In *2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. 1–5. doi:10.23919/SOFTCOM.2018.8555764
[35] Hansol Woo, Gerald K. LeTendre, Trang Pham-Shouse, and Yuhan Xiong. 2021. The use of social robots in classrooms: A review of field-based studies. *Educational Research Review* 33 (6 2021). doi:10.1016/j.edurev.2021.100388
[36] Cristina Zaga, Manja Lohse, Khiet P. Truong, and Vanessa Evers. 2015. The effect of a robot's social character on children's task engagement: Peer versus tutor. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9388 LNCS. Springer Verlag, 704–713. doi:10.1007/978-3-319-25554-5_70
[37] Chunpeng Zhai, Santoso Wibowo, and Lily D. Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11 (12 2024). Issue 1. doi:10.1186/s40561-024-00316-7
[38] Xinyue Zhang, Fangqing Zhu, Kun Wang, Guitao Cao, Yaofeng Xue, and Mingzhuo Liu. 2024. Bring the Intelligent Tutoring Robots to Education: A Systematic Literature Review. *IEEE Transactions on Learning Technologies* (2024). doi:10.1109/TLT.2024.3428366
[39] Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, and Heng Ji. 2025. The Law of Knowledge Overshadowing: Towards Understanding, Predicting, and Preventing LLM Hallucination. (2 2025). http://arxiv.org/abs/2502.16143

Received 2025-09-30; accepted 2025-12-01